How Should Financial Intermediation Services be Taxed?

Ben Lockwood

No 948

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

THE UNIVERSITY OF WARWICK

How Should Financial Intermediation Services be Taxed?*

Ben Lockwood†

First version: May 31 2010
This version: October 23 2010

Abstract

This paper considers the optimal taxation of savings intermediation and payment services in a dynamic general equilibrium setting, when the government can also use consumption and income taxes. When payment services are used in strict proportion to final consumption, and the cost of intermediation services is fixed and the same across firms, the optimal taxes are generally indeterminate. But, when firms differ exogenously in the cost of intermediation services, the tax on savings intermediation should be zero. Also, when household time and payment services are substitutes in transactions, the optimal tax rate on payment services is determined by the returns to scale in the conditional demand for payment services, and is generally different to the optimal rate on consumption goods. In particular, with constant returns to scale, payment services should be untaxed. These results can be understood as applications of the Diamond-Mirrlees production efficiency theorem. Finally, as an extension, we endogenize intermediation, in the form of monitoring, and show that it may be oversupplied in equilibrium when banks have monopoly power, justifying a Pigouvian tax in this case.

JEL Classification: G21, H21, H25

Keywords: financial intermediation services, tax design, banks, monitoring,payment services

*I would like to thank Steve Bond, Clemens Fuest, Michael Devereux, Michael McMahon, Miltos Makris, and participants at the 2010 CBT Summer Symposium for very helpful comments on an earlier draft. I also gratefully ackowledge support from the ESRC grant RES-060-25-0033, "Business, Tax and Welfare".

†CBT, CEPR and Department of Economics, University of Warwick, Coventry CV4 7AL, England; Email: B.Lockwood@warwick.ac.uk

1. Introduction

Financial intermediation services include such important services as intermediation between borrowers and lenders, insurance, and payment services (e.g. credit and debit card services). These services comprise a significant and growing part of the national economy; for example, financial intermediation services, measured using the OECD methodology[1], were 3.9% of GDP in the UK in 1970, and increased to 7.9% by 2005. The figures for the Eurozone countries as a whole are 2.7% to 5.5%. In the US, the finance and insurance sector, excluding real estate, which includes financial intermediation, accounted for 7.3% of US value-added in 1999, rising to 8.4% in 2009[2].

The question of whether, and how, financial intermediation services should be taxed is a contentious one. In the tax policy literature, it is largely assumed that within a consumption tax system, such as a VAT, it is desirable to tax financial services. For example, the European Commission has recently proposed changes to the VAT treatment of financial services within the European Union, so as bring these more within the scope of VAT (de la Feria and Lockwood (2010)). Also, the recent IMF proposals for a "bank tax" to cover the cost of government interventions in the banking system include a Financial Activities Tax levied on bank profits and remuneration, which would work very much like a VAT, levied using the addition method (IMF(2010)).

But, it is also recognized that there are technical difficulties in taxing financial intermediation when those services are not explicitly priced (so-called margin-based services), such as the intermediation between borrowers and lenders. This raises a problem for the use of a VAT via the usual invoice-credit method, for example (Ebril, Keen, Bodin and Summers(2001)). As a result of this, the status quo in most countries is that a wide range of financial intermediation services are not taxed[3]. However, conceptually, the problems can be solved, for example, by use of a cash-flow VAT (Hoffman et. al.(1987), Poddar and English(1997), Huizinga(2002), Zee(2005)), and the increasing sophistication of banks' IT systems means that these solutions are also becoming practical.

So, it is increasingly relevant to ask, setting aside the technical problems, should financial intermediation services supplied to households be taxed at all? And if so, at

[1]See http://www.euklems.net.

[2]See http://www.bea.gov/industry/gdpbyind_data.htm

[3]For example, in the EU, the Sixth VAT Directive and subsequent legislation exempts a wide range of financial services from VAT, including insurance and reinsurance transactions, the granting and the negotiation of credit, transactions concerning deposit and current accounts, payments, transfers, debts, cheques,currency, bank notes and coins used as legal tender etc. (Council Direective 2006/112/EC of 28 November 2006, Article 135).

what rates? Given the overall importance of financial services to modern economies, there is surprisingly little written on this more fundamental question (see Section 2 for a discussion of the literature). Moreover, we would argue that the existing literature does not really clarify which of the fundamental principles of tax design apply. For example, is it the case that financial intermediation services are intermediate goods in the production of final consumption for households, and thus should not be taxed? Or, should they be taxed at the same rate as other goods purchased on the market, at least under conditions when a uniform consumption tax is optimal?

The objective of this paper is to address these fundamental questions[4]. We set up and solve the tax design problem in a dynamic general equilibrium model of the Chamley(1986) type, where the government chooses taxes on payment services and savings intermediation, as well as the usual taxes on consumption (or equivalently, wage income) and income from capital, and where financial intermediaries, in the form of banks, are explicitly modelled. On the payment services side, we assume, following the literature on the transactions cost approach to the demand for money, that payment services are not necessarily proportional to consumption, but can be used to economize on the household time input to trading. This is realistic: for example, making use of a basic bank account requires a time input, e.g. trips to the bank, but use of an additional payment service e.g. a credit card, substitutes for trips to the bank.

We assume initially that the cost of savings intermediation per unit of capital is fixed, but can vary across borrowers (firms). Again, this is realistic; savings intermediation is a complex process involving initial assessment of the borrower via e.g. credit scoring, structuring and pricing the loan, and monitoring compliance with loan covenants (Gup and Kolari(2005, chapter 9). There is evidence that other things equal, the cost of borrowing is lower for firms that have had longer relationships with banks (Berger and Udell(1995)), or make information available to banks via rating agencies Brown, Jappelli and Pagano(2009)).

We then solve the tax design problem, where the government has access to a full set of taxes, i.e. the usual wage and capital income taxes, plus a tax on the consumption good and on payment services, and a tax paid the bank on the spread between borrowing and lending rates. In this set-up, there is the usual tax indeterminacy, as a uniform tax on payment services and the consumption good is equivalent to a wage tax. We set the wage tax equal to zero, and then show that in the general model, all remaining four taxes

[4]It should be noted that this paper does *not* deal with corrective taxes on bank lending designed to internalize the social costs of bank failure or the costs of bailout; on this, see e.g. Keen(2010).

are determinate i.e. there is no redundancy in tax instruments. However, under certain restrictive conditions - in fact, those assumed by the existing literature on this topic, see Section 2- the tax structure is indeterminate.

The tax on savings intermediation is determined as follows. In the tax design problem, the tax on capital income is used as the "instrument" to pin down the rate of substitution between present and future consumption for the household. So, this means that the tax on savings intermediation is a "free instrument" that can be used to ensure that capital is allocated efficiently across firms. In turn, the cost of capital to a particular firm will be the cost of capital to the bank i.e. the return paid to depositors, plus the cost of intermediation, where the latter includes any tax. So, a non-zero tax on savings intermediation will distort the relative cost of capital across firms, and so this tax is optimally set to zero. This is a version of the Diamond-Mirrlees production efficiency result.

Turning to the tax on the payment service, first, our first result is that the total tax "wedge" between consumption and leisure is a weighted average of the tax on consumption and on the payment service, and is determined by a standard optimal tax formula, involving the general equilibrium expenditure elasticity of consumption (Atkeson, Chari, and Kehoe(1999)). The sign of the tax on the payment service is itself determined[5] *not* by the structure of preferences, but by the returns to scale in the conditional demand for payment services as a function of the household consumption level and time input to transactions.

Specifically, with constant returns, payment services should be untaxed; this can be understood as an instance of the Diamond-Mirrlees production efficiency result. More generally, under reasonable assumptions, notably that there is a fixed element to transactions costs, there are decreasing returns to scale, and in this case, it is shown that payment services should be taxed. The intuition is somewhat similar to the Corlett-Hague rule; with decreasing returns, the household makes positive a notional "profit" from the production of the final consumption good using time and payment services. Given that this profit cannot be taxed directly, it is optimal to tax it indirectly via taxing one of the inputs, payment services. The general conclusion is that the tax on payment services is determined in a completely different way to the tax on consumption, and thus will in general be at a different rate.

Finally, in Section 5 of the paper, we consider how the savings intermediation tax

[5]Strictly speaking, this requires the conditional demand for payment services to be Cobb-Douglas, but it is also likely to hold for a variety of other cases, see Section 4.

should be designed when financial intermediation by banks is modelled explicitly. It is clear that banks supply several different kinds of intermediation services[6], notably liquidity services to households (Diamond and Dybvig(1983)), and monitoring services to firms (Diamond (1991), Besanko and Kanatas(1993), Holmstrom and Tirole (1997)). We argue that as long as these services are provided efficiently, i.e. there are no "market failures" in provision of intermediation services, more explicit modelling of them will not change the basic conclusions. If there *are* market failures, then these can be remedied by Pigouvian taxes, but these are *in addition* to the optimal tax structure identified in this paper[7].

We make this point by extending our model to allow for an endogenous amount of intermediation services (per unit of savings) in the from of monitoring, along the lines of Holmstrom and Tirole(1997)[8]. In their framework, without monitoring, bank lending to firms is impossible, because the informational rent they demand is so high that the residual return to the bank does not cover the cost of capital. So, as monitoring is costly, the socially efficient level of monitoring is that level which just induces to bank to lend. In the case where the bank is competitive, i.e. where firm chooses the terms of the loan contract subject to a break-even constraint for the bank, an assumption commonly made in the finance literature, this is also the equilibrium level of monitoring. In this case, savings intermediation should *not* be taxed, because doing to will violate production efficiency, as in the case with heterogenous firms and a fixed amount of intermediation services per unit of savings. But, in the case where the bank is a monopolist i.e. it chooses the contract, it will generally choose a *higher* level of monitoring than this, in order to reduce the firm's informational rent. So, in this case, the optimal tax is a positive Pigouvian tax, set to internalize this negative externality.

The remainder of the paper is organized as follows. Section 2 discusses related literature. Section 3 outlines the model, and explains how existing contributions can be viewed as special cases. Section 4 presents the main results. Section 5 studies the case of endogenous monitoring, and Section 6 concludes.

[6]See Swank(1996) for an overview of the differnt types of banking services.

[7]The same argument applies if banks engage in "gambling" with deposits (Hellmann, Murdock and Stiglitz(2000), Keen(2010), Miller, Zhang, and Li(2010)).

[8]It would, perhaps, be more natural to "endogenize" intermediation by looking at the provision of liquidity services using Diamond-Dybvig model, which is undoubtedly the pre-eminent microeconomic model of banking. While this is a topic for future work, the problem is that the Diamond-Dybvig model has a three-period dynamic structure, which is very difficult to embed within the standard infinite-horizon dynamic optimal tax model. In contrast, the Holmstrom-Tirole model also describes an important aspect of financial intermediation, and is essentially static, and can be embedded in this way.

2. Related Literature

There is a small literature directly addressing the of optimal taxation of borrower-lender intermediation and payment services, Grubert and Mackie(1999), Jack(1999), Auerbach and Gordon (2002), and Boadway and Keen(2003). Using for the most part a simple two-period consumption-savings model, these papers broadly agree on a policy prescription[9]. Given a consumption tax that is uniform across goods (at a point in time, or across time), payment services should be taxed at this uniform rate, but savings intermediation should be left untaxed. The argument used to establish this is simple; in a two-period consumption-savings model with the same, exogenously fixed, tax on consumption in both periods, this arrangement leaves the marginal rate of substitution between current and future consumption undistorted i.e. equal to the marginal rate of transformation[10].

However, one can make three criticisms of the current literature. First, even taking their set-up as given, their optimal taxes are indeterminate. Purely mathematically, two taxes cannot be uniquely determined from a single efficiency condition. Second, in their analysis, consumption (wage) and capital income taxes are taken as given, and not optimized by the government. Third, relative to the model of this paper, the models analyzed in the current literature are very special in a number of respects. For example, implicitly, these papers are assuming[11] a fixed labour supply, so that a uniform tax on consumption over the life-cycle is first-best efficient, as it does not distort the inter-temporal allocation of consumption, and thus financial intermediation should not do so either. Again, special assumptions are made about the demand for payment services, and intermediation activities of banks. Specifically, they assume (i) that payment services are consumed in proportion to consumption; (ii) that the costs of savings intermediation are in proportion to capital invested. We are able to show that the basic result of this literature - i.e. that intermediation taxes are indeterminate, but that *an* optimal tax structure is to tax payment services at the same rate as consumption, but exempt savings

[9] Chia and Whalley(1999), using a computational approach, reach the rather different conclusion that no intermediation services should be taxed, but but their model is not directly comparable to these, as the intermediation costs are assemed to be proportional to the *price* of the goods being transacted.

[10] Auerbach and Gordon have a model that is in some respects more general, and they also take a different analytical approach. Specifically, there model allows for T periods, multiple consumption goods, and variable labour supply. In this setting, they show that a uniform tax on all commodities and payment services is equivalent to a wage tax. Thus, they show that *if* a uniform commodity tax is optimal, payment services should be taxed at the same rate, consistently with the other literature cited.

[11] The exception here is Auerbach and Gordon(2002), where labour supply is variable. However, in their model, the consumption tax is just assumed to be uniform, not optimised.

intermediation - also emerges in our model when all of these special assumptions are made (Proposition 1 below).

A less closely related literature is that on the optimal inflation tax which take a transaction costs approach to the demand for money (Kimbrough(1986), Faig(1988), Guidotti and Vegh(1993), Correia and Teles(1999)). In this literature, money formally plays a role similar to payment services in our model; the main differences are (i) that it is assumed a free good i.e. it has a zero production cost, and (ii) it is subject to an inflation tax, rather than a fiscal tax. While (ii) makes no difference from an analytical point of view, (i) does; it turns out that when money is free, the optimal inflation tax is zero, as long as the transactions demand for household time is a homogenous function of money and consumption. A much more closely related finding is in Correia and Teles (1996), where, in Section 3 of their paper, money is allowed to have a positive production cost. Proposition 6 below can be regarded as an extension of Proposition 3 in their paper.

3. The Model

3.1. Households

The model is a version of Atkeson, Chari and Kehoe(1999) with payment services and savings intermediation. There is a single infinitely lived household with preferences over levels of a single consumption good, leisure, and a public good in each period $t = 0, ..\infty$ of the form

$$\sum_{t=0}^{\infty} \beta^t (u(c_t, l_t) + v(g_t)) \tag{3.1}$$

where c_t is the level of final consumption in period t, l_t is the consumption of leisure, and g_t is public good provision. Utilities $u(c, l), v(g)$ are strictly increasing and concave in their arguments.

We take a transactions cost approach to the demand for payment services[12], and suppose that consumption c_t incurs a transaction cost in terms of household time, and this cost is reduced by payment services x_t. For example, making use of credit requires a time input, e.g. trips to the bank, but use of an additional payment service e.g. a credit card substitutes for trips to the bank. Then we have $h_t = h(c_t, x_t)$, where h is increasing in c_t, and decreasing in x_t.

It turns out that for our purposes, it is convenient to describe the implicit relationship $h_t = h(c_t, x_t)$ between c_t, h_t, and x_t in terms of the *conditional demand for payment*

[12]This is of course, analagous to the transactions cost theory of the demand for money.

services

$$x_t = \phi(c_t, h_t) \tag{3.2}$$

where ϕ is increasing in c_t, and decreasing in h_t. We will assume that ϕ is homogenous of some fixed degree $\kappa > 0$. Specification (3.2) is convenient because it nests the existing literature as a special case: this literature effectively assumes $\kappa = 1$ and ϕ independent of h_t i.e. $x_t = Ac_t,\ A > 0$. It also allows simpler tax formulae than working with $h_t = h(c_t, x_t)$.

The household thus supplies labour to the market of amount

$$m_t = 1 - l_t - h_t \tag{3.3}$$

where the total endowment of time per period is set at unity. In each period t, the household also saves k_{t+1} in units of the consumption good, and deposits it with a bank, who can then lend it on to firms who can use it as an input to production in the next period, after which they must repay the loan to the bank, who then in turn repays the household. So, in this model, capital effectively only lasts one period[13].

Finally, in any period t, c_t, x_t are purchased by the household inclusive of taxes, and the household also pays proportional taxes on labour and capital income. There is a degree of indeterminacy in these tax instruments, as a uniform tax on consumption and payment services at rate τ is equivalent to a wage tax at rate $\frac{\tau}{1+\tau}$. So we assume w.l.o.g. that the wage tax on c_t is zero and denote the taxes on c_t, x_t by τ_t^c, τ_t^x. We also assume for convenience that one unit of the consumption good can be transformed into one unit of payment services or one unit of the public good. Moreover, in equilibrium, payment services are priced at marginal cost (see Section 3.3 below). This fixes the relative pre-tax price of z_t, x_t and g_t at unity.

So, the present value budget constraint of the household is

$$\sum_{t=0}^{\infty} p_t(c_t(1+\tau_t^c) + k_{t+1} + x_t(1+\tau_t^x)) = \sum_{t=0}^{\infty} p_t(w_t(1-\tau_t)m_t + (1+\rho_t)k_t) \tag{3.4}$$

where p_t is the price of output in period t, ρ_t is the after-tax return on capital to the household, and w_t is the wage. We normalize by setting $p_0 = 1$ and assume for convenience that $k_0 = 0$ i.e. initial capital is zero[14]. Finally, $\rho_t = (1-\tau_t^r)r_t$, where r_t is the pre-tax return on capital, determined below, and τ_t^r is the capital income tax.

[13]Mathematically, this is equivalent to the usual definition of capital in the dynamic optimal tax model, with a depreciation rate of 100% . But, the interpretation is slightly different - here, households do not rent capital to firms.

[14]This simplifies the implementability constraint, and does not change anything of substance (see Atkeson, Chari and Kehoe(1999).

Substituting (3.2),(3.3) in (3.4) gives:

$$\sum_{t=0}^{\infty} p_t(c_t(1+\tau_t^c)+k_{t+1}+\phi(c_t,h_t)(1+\tau_t^x)) = \sum_{t=0}^{\infty} p_t(w_t(1-l_t-h_t)+(1+\rho_t)k_t) \quad (3.5)$$

The first-order conditions for a maximum of (3.1) subject to (3.5) with respect to $c_t,\ l_t, h_t, k_{t+1}$ respectively are:

$$\begin{aligned} \beta^t u_{ct} &= \lambda p_t(1+\tau_t^c+\phi_{ct}(1+\tau_t^x)) && (3.6)\\ \beta^t u_{lt} &= \lambda p_t w_t && (3.7)\\ -\phi_{ht}(1+\tau_t^x) &= w_t && (3.8)\\ p_t &= (1+\rho_{t+1})p_{t+1} && (3.9) \end{aligned}$$

where λ is the multiplier on (3.5), and we use (here and below) the notation that for any any function f and variables $x_t\,,\ y_t\,$, the partial derivative of f with respect to x_t is f_{xt}, the cross-derivative is $_{f_{xyt}}$ etc. Note that using this notation, the consumer price of final consumption is $p_t\left(1+\tau_t^c+\phi_{ct}(1+\tau_t^x)\right)$, a weighted sum of the prices facing the household of c_t, and x_t.

3.2. Firms

There are firms, $i=1,..n$. Firm i produces output from labour and capital via the constant returns production function $F^i(k_t^i, m_t^i)$, where k_t^i, m_t^i are capital and labour inputs. These firms are assumed to be perfectly competitive. But, they cannot purchase capital directly from households, but must borrow from banks. Moreover, we suppose that firms may differ in intermediation costs, as described in more detail in Section 3.3 below. So, firms face differences in the cost of capital i.e. firm i must repay $1+r_t^i$ per unit of capital borrowed from the bank. Thus, profit-maximization implies:

$$F_m(k_t^i, m_t^i) = w_t,\ \ F_k^i(k_t^i, m_t^i) = 1+r_t^i \quad (3.10)$$

And, in addition, the capital and labour market clearing conditions are:

$$\sum_{i=1}^{n} k_t^i = k_t,\ \sum_{i=1}^{n} m_t^i = 1-l_t-h_t \quad (3.11)$$

These conditions (3.10),(3.11) jointly determine w_t and r_t , given household savings and labour supply decisions.

3.3. Banks

Banks in this economy provide two possible services. First, they can provide payment services to the households i.e. supply x_t. Second, they can provide intermediation between households and firms. Banks can compete on price for both these activities (i.e. households see the banks as perfect substitutes, both with respect to payment and intermediation services). We also assume no economies of scope, and constant returns in the provision of both services, so that banks must break even on both services. Assuming w.l.o.g. that the marginal and average cost of payment services is 1 in units of the consumption good, the price of payment services will also be 1 in equilibrium.

The cost of intermediating one unit of savings between the household and firm i is s^i. Note that we take s^i as fixed, but possibly varying between firms, for reasons discussed in the introduction. We also suppose that "spread" i.e. the value of intermediation services provided by the bank, can be taxed at some rate $\tilde{\tau}_t^s$. In turn, the value of intermediation services is measured by $r_t^i - r_t$, where r_t^i is the lending rate to firm i, and r_t is the rate paid to depositors. So, $\hat{\tau}_t^s$ is a tax on both intermediation services provided to households, and to firms[15]. Then, as banks make zero profit on this activity, we must have

$$(1 - \hat{\tau}_t^s)(r_t^i - r_t\ ) - s^i = 0,\ i = 1, ..n \tag{3.12}$$

Then, from (3.12):

$$r_t^i = r_t\ + (1 + \tau_t^s)s^i, \quad \tau_t^s = \frac{\hat{\tau}_t^s}{1 - \hat{\tau}_t^s} \tag{3.13}$$

We refer to τ_t^s as the *spread tax* from now on.

3.4. Discussion

The above model provides a general framework which encompasses the specific models of taxation of financial services (Auerbach and Gordon(2002), Boadway and Keen(2003)), Jack(1999), Grubert and Mackie(1999)) that have been developed so far. For example, Boadway and Keen(2003)), Jack(1999), Grubert and Mackie(1999) are two-period versions of the above model[16], with (implicitly) fixed labour supply. Auerbach and Gordon(2002) is a finite-horizon version of the model, with the additional feature[17] that there are n

[15] In principle, one could allow for the intermediation services received by these two parties to be taxed at different rates, but in practice, this is very difficult to implement (Poddar and English(1997)).

[16] A minor qualification here is that Boadway and Keen allow for a fixed cost of savings intermediation e.g. fixed costs of opening a savings account. These introduce a non-convexity into household decision-making, which greatly complicates the optimal tax problem, and so we abstract from these in this paper.

[17] It also has labour supply in only one period.

consumption goods in each period, a feature that, however, is inessential in the sense that the main results of this paper generalize straightforwardly to n consumption goods in each period (see Section 5 below).

As already noted in Section 2, the feature of all these contributions, however, is the special assumptions they implicitly make about demand for payment services and bank intermediation. On the household side, they all assume, first, that payment services are needed *in fixed proportion to consumption* and that (implicitly) that a time input h_t is not required from the household. In our model, this amounts to the assumptions that $\phi(c_t, h_t) = Ac_t$ in (3.2), in which case, choosing the constant to be unity, $x_t = c_t$. On banking activity, the existing literature assumes that the cost of intermediation *in fixed proportion to household savings.* In the context of our model, this requires $s^i = s$ i.e. firms are all the same with respect to intermediation costs, or - equivalently - there is only one firm.

Finally, the relation of our model to the optimal inflation tax literature is as follows. Our modelling of household demand for intermediation services is closely related to the "transactions cost" view of the demand for money in that literature (Corriea and Teles(1996), (1999)). In particular, if we define x_t as real money balances, their transactions cost function is and inversion of (3.2) to obtain h_t as a function of c_t, x_t; then, increased real money balances reduce the labour transactions costs of consumption. The models in this literature do not allow for physical capital or taxation of capital income, or costly money, and so in this sense are more special. Nevertheless, one of our results, Proposition 3 below, is related to that literature, especially Proposition 2 of Correia and Teles(1996).

3.5. A Benchmark Indeterminacy Result

Now we make the assumptions of the existing literature (Auerbach and Gordon(2002), Boadway and Keen(2003)), Jack(1999), Grubert and Mackie(1999)), namely: (i) that conditional demand for x_t is independent of h_t and linear in c_t i.e. $x_t = Ac_t$; (ii) only one type of firm; and (iii) a fixed consumption tax τ_t^c and a zero capital income tax, $\tau_t^r = 0$. Under these assumptions, we show that optimal taxes on financial intermediation are generally indeterminate. Note from (3.6)-(3.9) that given (i) i.e. $\phi_{ct} = 1$, and $\tau_t^r = 0$, we have:

$$\frac{\beta u_{ct}}{u_{ct-1}} = \frac{1 + \tau_t^c + A(1 + \tau_t^x)}{1 + \tau_{t-1}^c + A(1 + \tau_{t-1}^x)} \frac{1}{1 + r_t}, \; t = 1, .. \tag{3.14}$$

Moreover, from (3.10), (3.13), given only one firm:

$$r_t = F_{kt} - 1 - (1 + \tau_t^s)s, \ t = 1, .. \tag{3.15}$$

where $F_{kt} = F_k(m_t, k_t)$. Then (3.14) becomes

$$\frac{\beta u_{ct}}{u_{ct-1}} = \frac{1 + \tau_t^c + A(1 + \tau_t^x)}{1 + \tau_{t-1}^c + A(1 + \tau_{t-1}^x)} \frac{1}{F_{kt} - (1 + \tau_t^s)s}, \ t = 1, .. \tag{3.16}$$

Now say that the sequence $\{\tau_t^x, \tau_t^s\}_{t=0}^{\infty}$ is a *restricted optimal tax structure on financial services* if the inter-temporal allocation of consumption is left undistorted by taxes. From (3.16), this requires:

$$\frac{1 + \tau_t^c + A(1 + \tau_t^x)}{1 + \tau_{t-1}^c + A(1 + \tau_{t-1}^x)} . \frac{1}{F_{kt} - (1 + \tau_t^s)s} = \frac{1}{F_{kt} - s}, \ t = 1, .. \tag{3.17}$$

Then two conclusions that can easily be drawn from (3.17). First, $\{\tau_t^x, \tau_t^s\}_{t=0}^{\infty}$ is not uniquely determined from (3.17) i.e. there is indeterminacy in the restricted optimal tax structure. The second is that of the many optimal tax combinations, $\tau_t^x = \tau_t^c, \ \tau_t^s = 0$ has the advantage that it is optimal, independently of knowledge of F_k, s and is thus administratively convenient. We can thus summarize:

Proposition 1. *In the benchmark case, with (i) conditional demand for x_t independent of h_t and linear in c_t; (ii) only one type of firm; and (iii) a fixed consumption tax τ_t^c and zero capital income tax, $\tau_t^r = 0$, then the restricted optimal tax structure on financial services is not uniquely determined. But, a uniform tax on goods and payment services ($\tau_t^x = \tau_t^c$), and a zero tax on the spread ($\tau_t^s = 0$) is an administratively convenient restricted optimal tax structure.*

This result summarizes the findings of the existing literature, in the context of our model. It is important to emphasize that under the assumptions made by the existing literature, optimal taxes on financial intermediation are in fact *indeterminate*. This main purpose of this paper is to relax these assumptions in an empirically plausible way, and at the same time generate determinacy in the tax structure.

4. Tax Design

We take a primal approach to the tax design problem. In this approach, a policy for the government is a choice of all the primal variables in the model, in this case $\left\{c_t, l_t, h_t, k_{t+1}, g_t, (k_t^i, m_t^i)_{i=1}^n\right\}_{t=0}^{\infty}$ to maximize utility (3.1) subject to the capital and labour

market clearing conditions (3.11), aggregate resource, and implementability constraints. We are thus assuming, following Chamley(1986), that the government can pre-commit to policy at $t = 0$. The aggregate resource constraint says that total production must equal to the sum of the uses to which that production is put:

$$c_t + \phi(c_t, h_t) + k_{t+1} + g_t + \sum_{i=1}^{n} s^i k_t^i = \sum_{i=1}^{n} F^i(k_t^i, m_t^i), \quad t = 0, 1, .. \tag{4.1}$$

The implementability constraint ensures that the government's choices also solve the household optimization problem. First, using the fact from (3.2) that ϕ has constant returns of degree κ , we have:

$$\phi = \phi_{ct} c_t + \phi_{ht} h_t + \frac{\kappa - 1}{\kappa} \phi \tag{4.2}$$

Substituting (4.2) back into (3.5), we obtain:

$$\begin{aligned} & \sum_{t=0}^{\infty} p_t (c_t (1 + \tau_t^c + \phi_{ct}(1 + \tau_t^x)) + \phi_{ht}(1 + \tau_t^x) h_t + k_{t+1}) \qquad (4.3) \\ = & \sum_{t=0}^{\infty} p_t \left(w_t (1 - l_t - h_t) - \frac{(\kappa - 1)}{\kappa} \phi (1 + \tau_t^x) + (1 + \rho_t) k_t \right) \end{aligned}$$

Then, using the household's first-order conditions (3.6)-(3.9) in (4.3), we get the government's implementability constraint:

$$\sum_{t=0}^{\infty} \beta^t (u_{ct} c_t - (u_{lt}(1 - l_t) + u_{lt} \pi_t)) = 0 \tag{4.4}$$

where in (4.4), the expression:

$$\pi_t = -\frac{(1 - \kappa)}{\kappa} \frac{\phi(c_t, h_t)}{\phi_{ht}(c_t, h_t)} \tag{4.5}$$

is the notional "profit", in units of leisure, that the household makes from the activity of combining x_t and h_t to "produce" c_t. Note as $\phi_{ht} < 0$, $\pi_t > 0$ iff there are decreasing returns to scale in the conditional demand for x_t i.e. $\kappa < 1$.

So, following the primal approach, we can define the social welfare function

$$W_t = u(c_t, l_t) + v(g_t) + \mu \left(u_{ct} c_t - u_{lt}(1 - l_t + \pi_t) \right) \tag{4.6}$$

where $\mu \geq 0$ measures the cost of distortionary taxation. The government's choice of primal variables must maximize $\sum_{t=0}^{\infty} \beta^t W_t$ subject to (4.1) and (4.4). The first-order

conditions with respect to $c_t, l_t, h_t,\ k_t, g_t, k_{it}, m_{it}$ are, respectively;

$$\begin{aligned}
\beta^t W_{ct} &= (1+\phi_{ct})\zeta_t && (4.7)\\
\beta^t W_{lt} &= \zeta_t^m && (4.8)\\
-\beta^t \mu u_{lt}\pi_{ht} &= \zeta_t^m + \zeta_t \phi_{ht} && (4.9)\\
\zeta_t^k &= \zeta_{t-1} && (4.10)\\
\beta^t v_{gt} &= \zeta_t && (4.11)\\
\zeta_t F_{kt}^i &= \zeta_t^k + \zeta_t s^i && (4.12)\\
\zeta_t F_{mt}^i &= \zeta_t^m && (4.13)
\end{aligned}$$

where $\zeta_t,\ \zeta_t^k, \zeta_t^m$ are the multipliers on the resource, capital market, and labour market conditions at time t respectively.

Moreover, from (4.6),

$$\begin{aligned}
W_{lt} &= u_{lt}(1+\mu(1+H_{lt})), && (4.14)\\
H_{lt} &= \frac{-u_{llt}(1-l_t-\pi_t) - u_{lct}c_t}{u_{lt}}
\end{aligned}$$

and

$$\begin{aligned}
W_{ct} &= u_{ct}\,(1+\mu(1+H_{ct})), && (4.15)\\
H_{ct} &= \frac{u_{cct}c_t - u_{clt}(1-l_t+\pi_t) - u_{lt}\pi_{ct}}{u_{ct}}
\end{aligned}$$

So, H_{ct} is what Atkeson, Chari and Kehoe(1999) call the general equilibrium expenditure elasticity. Note that if there are constant returns to scale, i.e. $\kappa = 1,\ \pi_t \equiv 0$, and so H_{lt}, H_{ct} are reduce to standard formulae found, for example, in the primal approach to the static tax design problem (Atkinson and Stliglitz(1980)).

We begin by characterizing the overall tax on final consumption, which from (3.6) is the weighted sum of τ_t^c and τ_t^x i.e. $\tau_t^c + \phi_{ct}\tau_t^x$. We can then state (all proofs in Appendix):

Proposition 2. *At any date t, the optimal total tax on final consumption in ad valorem form is*

$$\frac{\tau_t^c + \phi_{ct}\tau_t^x}{1+\phi_{ct}+\tau_t^c+\phi_{ct}\tau_t^x} = \left(\frac{v_{gt}-\alpha_t}{v_{gt}}\right)\left(\frac{H_{lt}-H_{ct}}{1+H_{lt}}\right),\ \alpha_t = \frac{u_{lt}}{w_t} \qquad (4.16)$$

Note that (4.16) is a formula for an optimal consumption tax that also occurs in the static optimal tax problem, when the primal approach is used (Atkinson and Stiglitz(1980, p377). In particular, v_{gt} is the marginal benefit of \$1 to the government, and α_t is a measure of the marginal utility of \$1 to the household, so $\frac{v_{gt}-\alpha_t}{v_{gt}}$ is a measure of the

social gain from additional taxation at the margin. But, inspection of (4.14) and (4.15) reveals that in our analysis, the H_{lt}, H_{ct} are generally different to the static case, unless $\pi_t = 0$, which occurs when there are constant returns in the conditional demand for payment services, $\kappa = 1$. Note also that the optimal tax $\tau_t^c + \phi_{ct}\tau_t^x$ on final consumption is a weighted average of two taxes on marketed goods, c_t and x_t, and thus these two separate taxes are not yet determinate.

The next result characterizes τ_t^x, and can be stated as follows[18]:

Proposition 3. *If household demand for payment services depends on the time input* ($\phi_h < 0$), *any date* t, *the optimal ad valorem tax on payment services is*

$$\frac{\tau_t^x}{1+\tau_t^x} = -\frac{\mu\alpha_t}{v_{gt}}\pi_{ht} \tag{4.17}$$

where

$$\pi_{ht} = -\frac{(1-\kappa)}{\kappa}\left(\frac{\phi_t\phi_{hht}}{(\phi_{ht})^2} - 1\right) \tag{4.18}$$

is the marginal effect of h_t *on household profit (4.5). But, if conditional demand for payment services is independent of the time input* ($\phi_{ht} = 0$), *then the optimal tax on payment services is indeterminate.*

That is, generally, τ_t^x is determinate, but under the special conditions of the existing literature, when $\phi_{ht} = 0$, it is not. What can we say about the structure of optimal taxes in that special case? From (4.16), there are an infinite number of combinations of τ_t^c, τ_t^x that can be optimal. But, it is also clear from (4.16) that a total ad valorem tax on final consumption some percentage rate can be implemented by τ_t^c and τ_t^x set at the same percentage rate. This of course echoes Proposition 1, but is in fact a generalization of it, because we are considering the full, not the restricted, optimal tax problem.

Turning to the main case of interest, when τ_t^x *is* determinate, we see that it is not general equal to τ_t^c, but is instead determined by the effect of h_t on the the notional "profit" of the household, π_t. Specifically, the sign of τ_t^x is the sign of $-\pi_{ht}$. One intuition for this is as follows. If the government imposes a positive tax on x_t, this will cause a reduction in x_t, and at a fixed level of consumption, c_t, a compensating increase in h_t. If this decreases notional profit for the household, which is not directly taxable, this is

[18] As noted in Section 2, Proposition 3 is related to Proposition 2 of Corriea and Teles(1996). They consider what is formally a very similar tax design problem. The main differences are; (i) Proposition 6 extends their analysis by providing an explicit formula for the optimal tax rate, and characterizing the case where there are no labour transactions costs associated with consumption; (ii) they work with a different specification of (3.2), namely where h_t is the dependent variable.

desirable. But this last effect is measured just by $-\pi_{ht}$. Note that in the special case of constant returns, $\kappa = 1$, $\tau_t^x = 0$. This can be understood as an instance of the Diamond-Mirrlees Theorem; if household "profit" is zero, the intermediate good, payment services, should not be taxed.

More generally, there is an analogy here with the Corlett-Hague rule, which says that goods complementary with non-taxable leisure should be taxed more heavily. An analogy can also be drawn with tax design when there are non-constant returns to scale in the production of marketed goods. In that case, it has long been known that in this situation, a deviation from aggregate production efficiency (non-taxation of intermediate goods) is optimal. For example, Stiglitz and Dasgupta(1971) show that factors of production should be taxed more heavily when used in industries where pure rent is positive and cannot be taxed at 100%. Here, the principle is similar: the factor of production, x_t, should be taxed (subsidized), if it causes - indirectly, via h_t - profit to rise (fall).

We can now focus on the determinants of the sign of π_{ht}. We can start with the Cobb-Douglas case where (3.2), $\phi(c,h) = c^{\kappa+\theta}h^{-\theta},\ \theta \geq 0,\ \kappa > 0$. Then, $\frac{\phi_t\phi_{hht}}{(\phi_{ht})^2} = \frac{1}{\theta} + 1$, so that from (4.18), we see that

$$\frac{\tau_t^x}{1+\tau_t^x} = \frac{\mu\alpha_t}{v_{gt}}\frac{(1-\kappa)}{\theta\kappa} \tag{4.19}$$

so that sign of τ_t^x is determined by the returns to scale in conditional demand. In particular, if there are decreasing returns to scale, which is the plausible case, then τ_t^x is positive. However, it is possible for $\frac{\phi_t\phi_{hht}}{(\phi_{ht})^2} - 1$ to be negative, for example, if $\phi(c,h) = A\ln(c/h)c^{\kappa}$. Then $\frac{\phi_t\phi_{hht}}{(\phi_{ht})^2} - 1 = A\ln(c/h) - 1$, which could be negative for small enough A. So, the sign of the optimal payment tax is not always positive when there are decreasing returns in ϕ.

We now turn to the tax on capital income and the spread tax. Then, we have:

Proposition 4. *At any date t, the optimal taxes τ_t^r, τ_t^s satisfy*

$$\left(1 + (1-\tau_t^r)\left(\frac{\zeta_{t-1}}{\zeta_t} - 1 - \tau_t^s s^i\right)\right) = \frac{A_t}{A_{t-1}}\frac{\zeta_{t-1}}{\zeta_t},\ i = 1,..n \tag{4.20}$$

where $A_t = \frac{1+\phi_{ct}}{1+\mu(1+H_{ct})}$. So, if firms are homogenous in intermediation costs, ($s^i = s$, all i), then τ_t^r , τ_t^s are not uniquely determined, but if there is heterogeneity ($s^i \neq s^j$, some i,j) then the unique solution to the system (4.20) has $\tau_t^s = 0$, and in the steady state, $\tau_t^r = 0$.

So, we see that as long as intermediation costs differ across firms, the spread tax τ_t^s at any date should be zero. The intuition for this result is clear. From (4.12), (4.10), we

see that at any date t,

$$(F^i_{kt} - s^i) = \frac{\zeta_{t-1}}{\zeta_t} \Longrightarrow F^i_{kt} - s^i = F^j_{kt} - s^j \tag{4.21}$$

That is, the marginal product of capital net of true intermediation costs should be equal across firms, which of course is just the condition for capital to be allocated efficiently across firms. But, condition (4.21) is generally not consistent with a non-zero spread tax when firms are heterogenous, as then from (3.10), (3.13),

$$F^i_{kt} = 1 + r_t \ + (1 + \tau^s_t)s^i \Longrightarrow F^i_{kt} - s^i = 1 + r_t \ + \tau^s_t s^i$$

So, if $\tau^s_t s^i \neq \tau^s_t s^j$, (4.21) cannot hold. This is just an instance of the Diamond-Mirrlees production efficiency theorem. A tax on the spread is an intermediate tax on the allocation of capital, and given our assumptions (a full set of tax instruments, and no pure profits), this tax should be set to zero. Note also that when there is only one firm, this argument has no bite, and thus τ^s is left indeterminate.

Finally, we see that in the steady state, $\tau^r_t = 0$. So, the celebrated result of Chamley(1986) that in the steady state, the tax on capital income is zero continues to hold in our setting. In this sense, the optimal structure of wage and capital income taxes is separable from the optimal tax on borrower-lender intermediation.

5. Endogenizing Intermediation Services

We have, so far, treated the service of savings intermediation by banks in rather "black box" fashion. In particular, we have treated s, the amount of intermediation services per unit of capital, as exogenous. However, it is clear that banks supply several different kinds of intermediation services[19], notably liquidity services (Diamond and Dybvig(1983)), and monitoring services (Diamond (1991), Besanko and Kanatas(1993), Holmstrom and Tirole (1997)). In this section, we present a simple version of Holmstrom and Tirole (1997), where the role of banks is endogenous. Banks provide monitoring services, which enables them to reduce the informational rent of the firms to which they lend, and thus overcome the "credit rationing problem"[20] which prevents firms borrowing directly from households. Thus, in terms of the baseline model, we essentially endogenize s, the level

[19] See Swank(1996) for an overview of the differnt types of banking services.

[20] The credit rationing problem arises when the informational rent demanded by borrowers in exchange for "behaving" is so high that the residual return to the bank does not cover its cost of capital(Tirole(2006)).

of intermediation services provided per unit of capital intermediated. This micro-founded "sub-model" is then embedded in the general equilibrium model[21], which allows us to solve for the optimal tax on the spread, τ^s. We find that when the bank is "competitive" i.e. firms set the terms of the loan contract, $\tau^s = 0$, but that when the bank is a monopolist, τ^s is strictly positive.

The details are as follows (in what follows,we drop time subscripts except where necessary). First, we drop payment services from the model, by setting $x_t = h_t = 0$. We then assume that there are two kinds of firms, a non-entrepreneurial firm (NE-firm), and a continuum of unit measure of entrepreneurial firms (E-firms). The NE-firm has a linear production function $F(k, m) = wm + (1 + r)k$. Moreover, the NE-firm does not require bank intermediation, but can rent capital directly from the household. Given this linear production function, the cost of capital is determined independently of the capital stock at $1 + r$, and thus we require $\beta(\delta + r) = 1$ to ensure a steady state (Atkeson, Chari and Kehoe(1999)). If this holds, then the economy converges immediately to the steady state.

Each E-firm operates as follows. In any period t, it has a discrete investment project, which requires one unit of capital. The E-firm must borrow all of this; it has no collateral[22]. If the project is a success, it produces R units of output at $t+1$, and if it fails, it produces 0. Let $\tilde{R}$ be the random output of the project, assumed uncorrelated across E-firms. Following Holmstrom and Tirole(1997), Section 4.4, we model monitoring by supposing that there are two versions of the project that can be chosen by the firm: a good version, where the probability of success is $1 > q_H > 0$; and a bad version, where the probability of success is $0 < q_L = q_H - \Delta$, but there is a private benefit $b(s)$ for the E-firm, where $b(s)$ is further discussed below.

There is now a single bank, who lends to all the E-firms. The bank, as well as the E-firm, can observe $\tilde{R}$. Moreover, the bank has access to a monitoring technology, which operates as follows. If operated at intensity s, the bank can prevent the firm undertaking a bad project with payoff greater than $b(s)$, with $b(0) > 0,\ b' < 0,\ b'' > 0$. This can be interpreted, for example, as the extent to which the bank can detect whether the firm is conforming with the covenants of its loan contract. As in Holmstrom and Tirole(1997),

[21] It goes without saying that combining an agency model of banks with a dynamic tax design problem is a challenging exercise, and so we work with the simplest possible model of endogenous intermediation that we can formulate. In particular, we chose the Holmstrom-Tirole model to make our point because it seemed (to us) that it was not possible to embed the Diamond-Dybvig model in our dynamic model of optimal tax in a tractable way.

[22] This is in contrast to Holmstrom and Tirole, where the distribution of collateral across firms plays a central role.

we assume that only the good project is economically viable, even taking into account the maximum benefit $b(0)$ for the E-firm i.e.

$$q_H R - (1+r) > 0 > q_L R - (1+r) + b(0) \tag{5.1}$$

Following Holmstrom and Tirole(1997), section 4.4, we define a *loan contract* between the bank and the firm to be a repayment R_b from the firm to the bank, conditional on stochastic output $\tilde{R}$, with the firm retaining $R_e = \tilde{R} - R_b$, plus a level of monitoring, s. In the design of this contract, we assume limited liability, which means both payments R_b, R_e have to be non-negative. So, limited liability means that contract specifies $R_B = R_e = 0$ whenever output is zero.

The usual approach in the financial contracting literature is to study the optimal contract between the bank and the firm where the firm has all the "bargaining power" i.e. the bank is a passive entity which must simply make non-negative expected profit (Holmstrom and Tirole(1997), Tirole(2006)). We begin with this case as a benchmark, assuming that firms are risk-neutral, i.e. that they maximize expected profits[23].

From (5.1), the optimal contract must induce the firm to choose the good project i.e. it must satisfy the incentive constraint

$$q_H R_e \geq q_L R_e + b(s) \Longrightarrow R_e \geq \frac{b(s)}{\Delta} \equiv R_e(s) \tag{5.2}$$

It must also give the bank non-negative expected profit[24]. Generally, the profit of the bank can be written

$$\Pi^b(s, \tau^s) = (1 - \tau^s)\left(q_H(R - R_e) - (1+r)\right) - s \tag{5.3}$$

This is because $q_H(R - R_e) - (1+r)$ is the expected margin on the loan, or spread, which is taxed at rate τ^s, so (5.3) is of the same general form as (3.12) above. Assume for the moment that $\tau^s = 0$. Then the profit constraint of the firm is

$$\Pi^b(s, 0) = q_H(R - R_e) - (1+r) - s \geq 0 \tag{5.4}$$

The loan contract problem for the firm is then to choose R_e, s to maximize $q_H R_e$ subject to (5.2), (5.4). These constraints reduce to

$$q_H R - (1+r) - s \geq q_H R_e \geq q_H R_e(s) \tag{5.5}$$

[23]This can be justified as follows. There are a large number (a continuum) of firms, owned by the household, and because success probabilities are assumed uncorrelated, aggregate profit is non-stochastic and equal to the expected profit of a typical firm; so, the household prefers aggregate profit to be maximised.

[24]By the same aergument as in the previous footnote, we can assume that banks maximise expected profit.

Now assume:

A1. $q_H(R - R_e(0)) < (1 + r),\ \max_s \{q_H(R - R_e(s)) - s\} > (1 + r)$.

Assumption A1 is illustrated in Figure 1 below; note that $q_H(R - R_e(0))$ is concave as shown, because $b'' > 0$.

Figure 1 in here

The interpretation of A1 is as follows. The first inequality in A1 implies that without monitoring, lending to E-firms is impossible, because the informational rent they demand is so high that the residual return to the bank does not cover the cost of capital. In that case, we have credit rationing, in the terminology of Tirole (2006), Chapter 3. The second inequality in A1 ensures that, absent taxation, the credit rationing problem can be overcome by appropriate choice of monitoring.

So, from A1, (5.5) is violated at $s = 0$, but holds for s above a certain minimum level. So, it is clear that it is optimal for the firm to reduce s to the point where these inequalities *just* hold i.e. to $s = \underline{s}$, where $\underline{s}$ is the smallest root of

$$q_H(R - R_e(s)) - s = (1 + r),$$

as shown in Figure 1. Note that this is in fact the socially efficient level of monitoring; monitoring is costly, and so it is optimal to set it just at the point where the bank is willing to lend, and no higher.

What happens when the *bank* has all the bargaining however in designing the loan contract? Now the bank chooses R_e, s to maximize (5.3) subject to (5.2). Clearly, (5.2) will be binding, so substituting $R_e = R_e(s)$ into (5.3) and maximizing with respect to s, we get the first-order condition

$$-(1 - \tau^s) q_H R_e'(s) = 1 \Longrightarrow s = s^*(\tau^s) \tag{5.6}$$

for any fixed tax τ^s. Now from Figure 1, it is clear that $s^*(0) > \underline{s}$ i.e. without a tax, the bank "over-monitors" in order to reduce the informational rent of the E-firm. From (5.6), it is also clear that a tax $\underline{\tau^s}$ can be found which will make $s^*(\underline{\tau^s}) = \underline{s}$. But, at this tax, the firm will be making a negative profit, because

$$\begin{aligned} \max_s \Pi^b(s, \underline{\tau^s}) &= (1 - \underline{\tau^s}) \left(q_H(R - R_e(\underline{s})) - (1 + r)\right) - \underline{s} < \\ q_H(R - R_e(\underline{s})) - (1 + r) - \underline{s} &= 0 \end{aligned}$$

So, the best that the government can do it is to increase τ^s to the value $\tilde{\tau}^s$ at which $\max_s \Pi^b(s, \tilde{\tau}^s) = 0$; generally, $0 < \tilde{\tau}^s < \underline{\tau^s}$, implying a level of monitoring $\tilde{s} = s^*(\tilde{\tau}^s) > \underline{s}$, as shown on Figure 1[25]. Thus, $\tilde{\tau}^s$ is a *constrained Pigouvian tax;* it corrects over-monitoring as far as is possible while respecting the break-even constraint of the bank.

So far, we have considered the problem of choosing τ_t^s in isolation. The full optimal tax problem is then as follows. The government chooses the primal variables, now including s_t, i.e. $\{c_t, l_t, k_{t+1}, g_t, s_t\}_{t=0}^{\infty}$ to maximize utility (3.1) subject to the resource and implementability constraints, and the constraint that s_t must be achievable, given the tax instrument τ_t^s. Assuming the government can set a 100% profit tax on both the bank and the E-firm, the implementability constraint is just given by[26]

$$\sum_{t=0}^{\infty} \beta^t (u_{ct} c_t - u_{lt}(1 - l_t)) = 0 \tag{5.7}$$

Given the above analysis, the aggregate resource constraint says that total production must equal to the sum of the uses to which that production is put, given the control that the government has over s_t. Using the fact that $F(m_t, k_t) = w(1 - l_t) + (1 + r)k_t$, and that lending to E-firms occurs iff $s_t \geq \tilde{s}$, we can write this constraint as

$$c_t + k_{t+1} + g_t + s_t = \begin{cases} w(1 - l_t) + (1 + r)(k_t - 1) + q_H R, & s_t \geq \tilde{s} \\ w(1 - l_t) + (1 + r)k_t & s_t < \tilde{s} \end{cases} \quad t = 0, 1, .. \tag{5.8}$$

where $\tilde{s} = \underline{s}$ when the firm chooses the loan contract, and $\tilde{s} = s^*(\tilde{\tau}^s)$ when the bank chooses the loan contract. Note that in (5.8), output increases by $q_H R - (1 + r) > 0$ if E-firms are financed, but a cost $\tilde{s}$ must be incurred.

So, following the primal approach, we can define the social welfare function

$$W_t = u(c_t, l_t) + v(g_t) + \mu \left(u_{ct} c_t - u_{lt}(1 - l_t) \right)$$

where $\mu > 0$ as long as the tax on profits is insufficient to fund the cost of the public good, g_t. As before, the government's choice of primal variables maximizes $\sum_{t=0}^{\infty} \beta^t W_t$ subject

[25] Note that $\max_s \Pi^b(s, \tilde{\tau}^s) = 0$ at the point where the function $q_H(R - R_e(e)) - s/(1 - \tilde{\tau}^s)$ is tangent to $1 + r$; this curve is shown as the bold dotted line in Figure 1.

[26] Unlike in the baseline model, both bank and E-firm can make positive profits, with the bank making $\max_s \Pi^b(s, \tau^s)$ and the E-firm making $\Pi^e = q_H R_e$. But, if these are taxed at 100%, the household has no non-wage income, and thus the implementability constraint is the usual one.

to (5.8). The first-order conditions with respect to c_t, l_t, k_t, g_t are, respectively;

$$\begin{aligned} \beta^t W_{ct} &= \zeta_t && (5.9)\\ \beta^t W_{lt} &= w\zeta_t && (5.10)\\ -\zeta_{t-1} + (1+r)\zeta_t &= 0 && (5.11)\\ \beta^t v_{gt} &= \zeta_t && (5.12) \end{aligned}$$

where ζ_t is the multiplier on (5.8). As convergence to the steady state is immediate from $t = 1$ onwards, $W_{ct},\ W_{lt}, v_{gt}$ are independent of t, implying $\zeta_t = \beta\zeta_{t-1}$, so (5.11) holds by assumption of $(1+r)\beta = 1$. Also, the optimal choice of s_t is;

$$s_t = \begin{cases} \tilde{s}, & \text{iff } q_H R - (1+r) - \tilde{s} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

The optimal tax structure can now be characterized as follows. First, let

$$H_l = \frac{-u_{ll}(1-l_t) - u_{lc}c}{u_l},\ H_c = \frac{u_{cc}c - u_{cl}(1-l_t) - u_l\pi_c}{u_c}$$

Then we have:

Proposition 5. *Assume A1. Then, at every date t, the optimal consumption tax τ^c is given by*

$$\frac{\tau^c}{1+\tau^c} = \left(\frac{v_g - \alpha}{v_g}\right)\left(\frac{H_l - H_c}{1+H_l}\right),\ \alpha_t = \frac{u_l}{w} \qquad (5.13)$$

and the optimal tax on capital income is $\tau^r = 0$. The optimal spread tax τ^s is as follows. If the E-firm chooses the loan contract, $\tau^s = 0$. But, if the bank chooses the loan contract, $\tau^s = \tilde{\tau}^s > 0$. In both cases, in equilibrium, the bank always loans to the E-firms.

So, the properties of the consumption and capital income tax are the same as in the base case i.e. as in Propositions 2 and 4, taking into account the simplifications made elsewhere in the model in this Section. The main question of interest is how τ^s is set. In the case where the bank is "competitive", $\tau^s = 0$. This is because if the bank is competitive, it is supplying the efficient level of monitoring, $\underline{s}$, and imposition of a tax will distort this; s will have to rise from its efficient level to compensate the bank and make it just willing to lend. If, on the other hand, the bank is a monopolist, it "over-monitors" in order to reduce the informational rent of the E-firm, above the minimum level $\underline{s}$. This imposes a negative externality on the E-firm, and so τ^s is a Pigouvian tax which ensures that the bank internalizes the negative externality, subject to respecting the zero profit constraint.

6. Conclusions

This paper has considered the optimal taxation of two types of financial intermediation services (savings intermediation, and payment services) in a dynamic economy, when the government can also use wage and capital income taxes. When payment services are used in strict proportion to final consumption, and the cost of intermediation services is the same across firms, the optimal taxes on financial intermediation are generally indeterminate. But, when firms differ in the cost of intermediation services, the tax on savings intermediation should be zero. Also, when household time and payment services are substitutes in household "production" of final consumption, the optimal tax rate on payment services is determinate, and is generally different to the optimal rate on consumption goods. Finally, as an extension, we endogenized the cost of intermediation, and showed that intermediation services (monitoring) may be oversupplied in equilibrium when banks have monopoly power, justifying a Pigouvian tax in this case.

There are two obvious limitations of the analysis. The first is that the government is assumed to be able to precommit to a tax policy at time zero. However, even in a simpler setting without a banking sector, the characterization of the optimal time-consistent capital and labour taxes is a technically demanding exercise (see e.g. Phelan and Stacchetti (2001)) and so such an extension is certainly beyond the scope of this paper.

The second is the restriction to linear income taxation. The classic result of Atkinson and Stiglitz tells us that with non-linear income taxation, commodity taxation is redundant, and more recently, Golosov et. al. (2003) has recently shown that this result generalizes to a dynamic economy. Their result would apply, for example, in a version of our model where households differ in skill levels, and without any financial intermediation. What would happen if we introduced financial intermediation in this environment? The results on taxation of payment services seem likely to be affected, as the government has additional degrees of freedom with which to tax the notional "profit" from household production.

7. References

Atkeson, A., V.V.Chari, and P.J.Kehoe (1999), "Taxing capital income: a bad idea", *Federal Reserve bank of Minneapolis Quarterly Review*, 23, 3-17

Atkinson, A., and J.E.Stiglitz (1980), *Lectures on Public Economics*, MIT Press

Auerbach, A. and R.H.Gordon (2002), "Taxation of financial services under a VAT", *American Economic Review*, 92, Papers and Proceedings, 411-416

Berger, A.N. and G. F. Udell (1995), "Relationship Lending and Lines of Credit in Small Firm Finance", *The Journal of Business*, 68, 351-381

Besanko, D. and G.Kanatas (1993), "Credit market equilibrium with bank monitoring and moral hazard", *The Review of Financial Studies*, 6, 213-232

Boadway, R. and M.Keen (2003), "Theoretical perspectives on the taxation of capital and financial services", in Patrick Honahan (ed.) *The Taxation of Financial Intermediation* (World Bank and Oxford University Press)„ 31-80.

Brown, M., T.Jappelli, and M.Pagano (2009), "Information sharing and credit: Firm-level evidence from transition countries", *Journal of Financial Intermediation* 18, 151–172

Chamley, C., (1986), "Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives", *Econometrica*, 54, 607-622

Chia, N-C. and J.Whalley, (1999), "The tax treatment of financial intermediaries", *Journal of Money, Credit and Banking*, 31, 704-19

de la Feria, R., and B. Lockwood (2010), "Opting for Opting In? An Evaluation of the European Commission's Proposals for Reforming VAT on Financial Services", forthcoming, *Fiscal Studies*

Correia, I., and P.Teles (1999), "The optimal inflation tax" *Review of Economic Dynamics* 2, 325-346

Diamond, D.W. (1991), "Monitoring and reputation: the choice between bank loans and directly placed debt", *Journal of Political Economy*, 99, 689-721

Diamond, D.W and Dybvig, P.H, (1983), "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91, 401-19

Ebrill, L., M. Keen, J-P. Bodin, and V. Summers (2001), *The Modern VAT* (Washington D.C.: International Monetary Fund)

Faig, M. (1988), "Characterisations of Optimal Tax on Money When it Functions as a Medium of Exchange", *Journal of Monetary Economics* 22, 137-48.

Golosov, M., N.Kocherlakota, and A.Tsyvinski, (2003), "Optimal Indirect and Capital Taxation", *Review of Economic Studies*, , 70, 569-587

Grubert, H. and J.Mackie (1999), "Must financial services be taxed under a consumption tax?", *National Tax Journal* 53, 23-40

Guidotti, P.E. and C.A. Vegh (1993), "The optimal inflation tax when money reduces transactions costs", *Journal of Monetary Economics* 31, 189 205

Gup, B.E., and J.W.Kolari (2005), *Commercial banking: The Management of Risk*, 3rd ed. , Wiley

Hellmann, T. F., K.C. Murdock and J.E. Stiglitz (2000), "Liberalization, Moral Hazard in Banking, and Prudential Regulation: Are Capital Requirements Enough?" *American Economic Review*, 90, 147-165

Hoffman, L.A., S.N. Poddar and J. Whalley (1987), "Taxation of Banking Services Under a Consumption Type, Destination Basis VAT" *National Tax Journal* 40, 547-554

Huizinga, H. (2002), "Financial Services – VAT in Europe?" *Economic Policy*, October, 499-534

Holmstrom, B. and J.Tirole (1997), "Financial intermediation, loanable funds, and the real sector", *Quarterly Journal of Economics*, 112, 663-691

International Monetary Fund (2010), *A Fair and Substantial Contribution by the Financial Sector*, Final Report for the G20

Jack, W. (1999), "The treatment of financial services under a broad-based consumption tax", *National Tax Journal* 53, 841-51

Keen, M. (2010)," Taxing and Regulating Banks", unpublished paper, IMF

Kimbrough, K. (1986), The optimum quantity of money rule in the theory of public finance, *Journal of Monetary Economics* 18, 277-284.

Miller, M., L. Zhang, and H. H.Li (2010), "Riding for a fall: monopoly banking with hidden tail risk", unpublished paper, University of Warwick

Phelan, C. and E.Stacchetti (2001), "Sequential Equilibria in a Ramsey Tax Model," *Econometrica*, 69, 1491-1518

Poddar, S.N., and M. English (1997), "Taxation of Financial Services Under a Value-Added Tax: Applying the Cash-Flow Approach" *National Tax Journal* 50, 89-111

Stiglitz, J. E., and P. Dasgupta (1971), "Differential Taxation, Public Goods, and Economic Efficiency" *The Review of Economic Studies*, 38, 151-174.

Swank,J., (1996), "Theories of the banking firm: a review of the literature", *Bulletin of Economic Research*, 48, 173-207

Tirole, J., (2006), *The Theory of Corporate Finance*, MIT Press

Zee, H.H. (2005), "A New Approach to Taxing Financial Intermediation Services Under a Value-Added Tax" *National Tax Journal* 53(1), 77-92

A. Appendix

A.1. Proofs of Propositions

Proof of Proposition 2. From (4.7), (4.8), (4.13),(3.10), we have

$$\frac{W_{ct}}{W_{lt}} = \frac{u_{ct}}{u_{lt}} \frac{1+\mu(1+H_{ct})}{1+\mu(1+H_{lt})} = \frac{1+\phi_{ct}}{w_t} \tag{A.1}$$

And, from (3.6),(3.7):

$$\frac{u_{ct}}{u_{lt}} = \frac{1+\tau_t^c+\phi_{ct}(1+\tau_t^x)}{w_t} \tag{A.2}$$

So, combining (A.1), (A.2) we get:

$$\frac{\tau_t^c+\phi_{ct}\tau_t^x}{1+\tau_t^c+\phi_{ct}(1+\tau_t^x)} = \frac{\mu(H_{lt}-H_{ct})}{1+\mu(1+H_{lt})} \tag{A.3}$$

Also, from (4.8),(4.11),(4.13),(3.10) we have:

$$\begin{aligned} u_{lt}(1+\mu(1+H_{lt})) &= F_{mt}^i v_{gt} = w_t v_{gt} \\ &\Longrightarrow \quad \mu = \frac{1}{1+H_{lt}} \frac{v_{gt}-\alpha_t}{\alpha_t}, \quad \alpha_t = \frac{u_{lt}}{w_t} \end{aligned} \tag{A.4}$$

Combining (A.3),(A.4) to eliminate μ, and rearranging, we get (4.16) as required. □

Proof of Proposition 3. From (3.8), (3.10), we have

$$\frac{\tau_t^x}{1+\tau_t^x} = \frac{w_t+\phi_{ht}}{w_t} \tag{A.5}$$

And from (4.9), (4.13), we get:

$$-\frac{\beta^t}{\zeta_t}\mu\frac{u_{lt}}{w_t}\pi_{ht} = \frac{F_{mt}^i+\phi_{ht}}{w_t} = \frac{w_t+\phi_{ht}}{w_t} \tag{A.6}$$

But then, combining (A.5),(A.6) and using (4.11) and $\alpha_t = \frac{u_{lt}}{w_t}$ we get

$$-\frac{\mu\alpha_t}{v_{gt}}\pi_{ht} = \frac{\tau_t^x}{1+\tau_t^x}$$

as required. □

Proof of Proposition 4. From (4.7), (4.15), we get

$$\frac{\zeta_{t-1}}{\zeta_t} = \frac{1+\phi_{ct}}{1+\phi_{ct-1}} \frac{\beta^{t-1}W_{ct-1}}{\beta^t W_{ct}} = \frac{A_{t-1}}{A_t\beta}\frac{u_{ct-1}}{u_{ct}} \tag{A.7}$$

where $A_t = \frac{1+\phi_{ct}}{1+\mu(1+H_{ct})}$. Next, from from (4.12), (4.10),

$$F^i_{kt} - s^i = \frac{\zeta^k_t}{\zeta_t} = \frac{\zeta_{t-1}}{\zeta_t} \tag{A.8}$$

So, combining (A.7) and (A.8), we get

$$\frac{u_{ct-1}}{u_{ct}} = \frac{A_t}{A_{t-1}}\beta(F^i_{kt} - s^i) \tag{A.9}$$

Next, using (3.6), (3.9), $\rho_t = (1-\tau^r_t)r_t$, and (3.13), we get:

$$\frac{u_{ct-1}}{u_{ct}} = \beta\left(1 + (1-\tau^r_t)r_t\right) = \beta\left(1 + (1-\tau^r_t)\left(F^i_{kt} - 1 - (1+\tau^s_t)s^i\right)\right) \tag{A.10}$$

Combining (A.9), A.10), and eliminating $\frac{u_{ct}}{u_{ct+1}}$, we get that:

$$\left(1 + (1-\tau^r_t)\left(F^i_{kt} - 1 - (1+\tau^s_t)s^i\right)\right) = \frac{A_t}{A_{t-1}}(F^i_{kt} - s^i),\ i = 1,..n \tag{A.11}$$

Finally, using (A.8) to substitute $F^i_{kt} - s^i$ by $\frac{\zeta_{t-1}}{\zeta_t}$ in (A.11), we get (4.20) as required. If $n = 1$, (4.20) is a single condition and thus τ^r_t, τ^s_t are not uniquely determined. If $n > 1$, (4.20) comprises a system of $n > 1$ equations, and it is easy to verify that $\tau^r_t = 1 - \frac{\frac{A_t}{A_{t-1}}\frac{\zeta_{t-1}}{\zeta_t} - 1}{\frac{\zeta_{t-1}}{\zeta_t} - 1}$, $\tau^s_t = 0$ is the unique solution to this system. So, $\tau^r = 0$ is a solution in the steady state. □

Proof of Proposition 5. The first two parts of the proposition can be proved along the lines of Propositions 2 and 4. To prove the last part, all we have to prove is that the government wishes to induce lending by the bank i.e. from (5.8) that

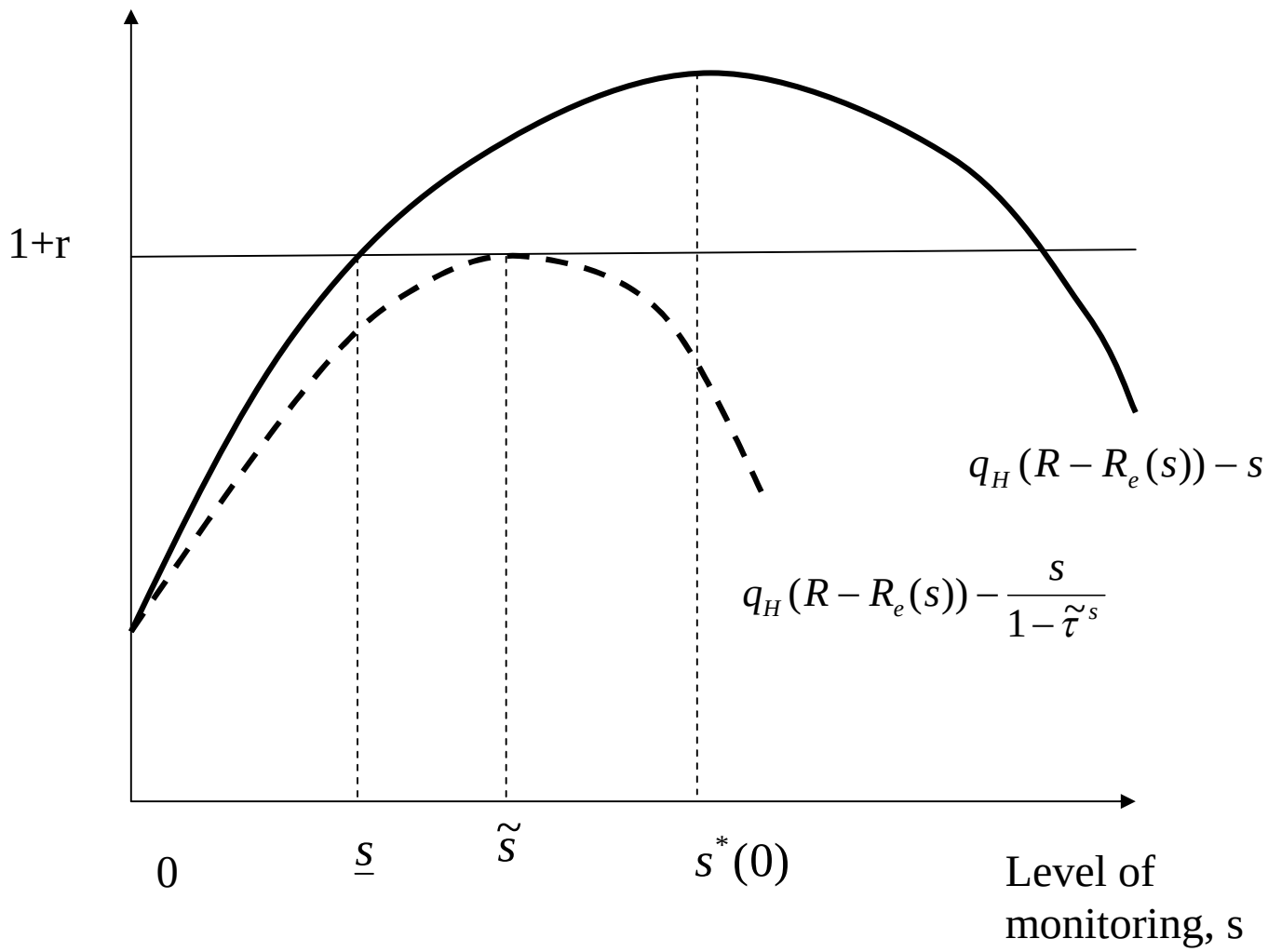
$$q_H R - (1+r) - \tilde{s} \geq 0$$

This condition is most stringent when $\tilde{s} = s^*(\tilde{\tau}^s)$. But, given the definitions of $\Pi^b(s,\ \tilde{\tau}^s), s^*(\tilde{\tau}^s)$, we see that

$$\begin{aligned} q_H R - (1+r) - \tilde{s} &> (1-\tilde{\tau}^s)(q_H R - (1+r)) - (1-\tilde{\tau}^s)q_H R_e(s^*(\tilde{\tau}^s)) - s^*(\tilde{\tau}^s) \\ &= \max_s \Pi^b(s,\ \tilde{\tau}^s) = 0 \end{aligned}$$

So, $s_t = \tilde{s}$ must be optimal. It then follows from the discussion in the text that the government must set $\tau^s = 0$ if the E-firm chooses the loan contract, and $\tau^s = \tilde{\tau}^s$ if the bank sets the loan contract. □

Figure 1


1+r
$q_H(R - R_e(s)) - s$
$q_H(R - R_e(s)) - \frac{s}{1-\tilde{\tau}^s}$
0
$\underline{s}$
$\tilde{s}$
$s^*(0)$
Level of monitoring, s